Metabarcoding and Metagenomics 5: 249–260
DOI 10.3897/mbmg.5.70063

Methods

Comparing species detection success between molecular markers in DNA metabarcoding of coastal macroinvertebrates

Barbara R. Leite[1,2,3], **Pedro E. Vieira**[1,2], **Jesús S. Troncoso**[3,4], **Filipe O. Costa**[1,2]

1 *Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*
2 *Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*
3 *UVIGO Marine Research Centre (CIM-UVIGO), ECIMAT Marine Station, Toralla Island, s/n, Vigo, Spain*
4 *Department of Ecology and Animal Biology, Marine Sciences Faculty, University of Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain*

Corresponding authors: Barbara R. Leite (barbaradrl.bio@gmail.com), Filipe O. Costa (fcosta@bio.uminho.pt)

Academic editor: Owen S. Wangensteen | **Received** 12 June 2021 | **Accepted** 7 December 2021 | **Published** 29 December 2021

Abstract

DNA metabarcoding has great potential to improve marine biomonitoring programs by providing a rapid and accurate assessment of species composition in zoobenthic communities. However, some methodological improvements are still required, especially regarding failed detections, primers efficiency and incompleteness of databases. Here we assessed the efficiency of two different marker loci (COI and 18S) and three primer pairs in marine species detection through DNA metabarcoding of the macrozoobenthic communities colonizing three types of artificial substrates (slate, PVC and granite), sampled between 3 and 15 months of deployment. To accurately compare detection success between markers, we also compared the representativeness of the detected species in public databases and revised the reliability of the taxonomic assignments. Globally, we recorded extensive complementarity in the species detected by each marker, with 69% of the species exclusively detected by either 18S or COI. Individually, each of the three primer pairs recovered, at most, 52% of all species detected on the samples, showing also different abilities to amplify specific taxonomic groups. Most of the detected species have reliable reference sequences in their respective databases (82% for COI and 72% for 18S), meaning that when a species was detected by one marker and not by the other, it was most likely due to faulty amplification, and not by lack of matching sequences in the database. Overall, results showed the impact of marker and primer applied on species detection ability and indicated that, currently, if only a single marker or primer pair is employed in marine zoobenthos metabarcoding, a fair portion of the diversity may be overlooked.

Key Words

COI, 18S, DNA metabarcoding, marine macrozoobenthic diversity, primer efficiency, taxonomic discrimination

Introduction

DNA metabarcoding is the taxonomic identification of organisms present in a bulk or environmental sample through the use of DNA amplification of standard regions of a genome (i.e. DNA barcodes) coupled with high-throughput sequencing (HTS) (Taberlet et al. 2012). DNA metabarcoding studies have been developed for diverse taxonomic groups (e.g. terrestrial arthropods: Elbrecht et al. 2019; freshwater macroinvertebrates: Bista et al. 2018; Giebner et al. 2020; meiofaunal organisms: Fais et al. 2020b; marine communities: Leray and Knowlton 2015; Aylagas et al. 2018), using a wide range of laboratory procedures (Andújar et al. 2018) and addressing questions about species richness, taxonomic composition, as well as biodiversity patterns (McGee et al. 2019; Piñol et al. 2019).

Metabarcoding allows for comparison across studies, however the harmonization and standardization of protocols is still far from being established (van der Loos and Nijland 2020; Duarte et al. 2021). While DNA-based approaches for assessing and monitoring marine

Copyright *Barbara R. Leite et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

macroinvertebrate species are constantly evolving (Andújar et al. 2018), the diversity of the adopted methodologies, including the use of different primer pairs or molecular markers, the lack of accurate and complete reference databases, and the continuous emergence of new sequencing innovations and bioinformatics pipelines, implies low standardization and comparability among studies (Cowart et al. 2015; Leray and Knowlton 2017), which remain important challenges that should be addressed.

Targeting marine species is especially challenging due to the broad taxonomic and phylogenetic composition of marine communities, and the choice of marker usually depends on the target taxa. However, the balance between the range of taxonomic coverage and the taxonomic discrimination ability should be considered in the choice of target genomic region and/or primer pairs, since it may affect the number of species and the taxonomic groups detected (van der Loos and Nijland 2020). Initially, a fundamental and critical decision in a DNA metabarcoding study is centered on which genomic region should be targeted. The 5' end of the mitochondrial cytochrome oxidase I gene (COI) is the standard barcode for animal life and the backbone of the universal Barcode of Life species identification system (Hebert et al. 2003), being the recommend marker for community metabarcoding (Andújar et al. 2018) and by far the most well represented genomic region in public databases (Porter and Hajibabaei 2020). After COI, the nuclear small subunit rRNA gene (18S) is among the most widely used markers in marine biodiversity studies (e.g. Lejzerowicz et al. 2015; Wangensteen et al. 2018; Zhang et al. 2018; Fais et al. 2020a; Duarte et al. 2021).

PCR-based methodologies are highly influenced by amplification biases thereby encouraging the use of several primer pairs in different metabarcoding studies (Bista et al. 2018; Elbrecht et al. 2019; Hajibabaei et al. 2019; Porter and Hajibabaei 2020). Amplicon length, primer mismatches, GC content, and polymerase errors can also affect the ability to detect the species present in bulk or environmental samples, (Kebschull and Zador 2015; Piñol et al. 2015; Nichols et al. 2018; Derycke et al. 2020). Even broad-range primers (i.e. able to amplify a DNA fragment over a broad taxonomic scope) demonstrated more affinity for some species and consequently do not perfectly match the DNA of all species present in a bulk sample. Low species detection ability has been associated with the use of non-degenerate primers (Clarke et al. 2014; Elbrecht et al. 2017; Collins et al. 2019), however, degenerate primers seem to have a higher capability to detect species compared with nondegenerate ones. The Leray-Geller fragment (mlCOIintF/jgHCO2198 – Leray et al. 2013, Geller et al. 2013) is a degenerate primer pair widely used in DNA metabarcoding studies targeting different types of taxa (Leray and Knowlton 2015; Clarke et al. 2017; Ransome et al. 2017; Aylagas et al. 2018), mostly due to their design for marine organisms with a wide phylogenetic coverage and fair amplicon length (313 bp). Yet, the inclusion of inosine bases in the reverse primer jgHCO2198 may impair the performance of the most commonly used high-fidelity polymerases for generating amplicon libraries for HTS (Knittel and Picard 1993, Jungbluth et al. 2021). The combination of mlCOIintF with LoboR1 amplifies exactly the same fragment and with demonstrated success in the amplification of DNA barcodes of marine taxa (Haenel et al. 2017; Hollatz et al. 2017; Chang et al. 2020; Castro et al. 2021). Primers amplifying the nuclear genes (e.g. TAReuk454FWD1/TAReukREV3, Lejzerowicz et al. 2015) constitute alternatives to COI primers, mainly because of their slower rate of evolution which results in more conserved regions which may facilitate the design of primers. However, reference databases for 18S target regions are less populated than COI (Andújar et al. 2018) and COI-primers often outperform primers for rDNA loci on taxon recovery (Clarke et al. 2017; Elbrecht et al. 2017, 2019; Atienza et al. 2020) and species discrimination ability (Tang et al. 2012; Clarke et al. 2017). Choosing the appropriate molecular markers and/or primer pairs is a principal step to influence the quality of data through PCR amplification, to accurately and efficiently determine the taxonomic composition of the bulk sample and to assign a species level identification.

For DNA metabarcoding studies, multiple sets of primers amplifying different molecular markers have been used to target a broad range of taxonomic groups in different marine communities. However, the majority of studies used a single primer pair or single marker loci strategy (Duarte et al. 2021). In two studies which comprehensively review DNA metabarcoding studies over the last 10 years, the authors concluded that only a minor portion of the publications used more than one marker (25% in van der Loos and Nijland 2020; 24% in Duarte et al. 2021).

Although DNA metabarcoding studies aim to provide species level assignments (Taberlet et al. 2012), the existence of gaps in the reference sequence databases (Weigand et al. 2019; Leite et al. 2020; Mugnai et al. 2021), associated with the limited species level discrimination ability of some markers (e.g. 18S rRNA; von Ammon et al. 2018) in some taxonomic groups, reduces the discrimination level of taxonomic identifications. However, the trade-off between the range of taxonomic coverage (species successfully amplified) and the taxonomic discrimination ability should be considered to enable accurate identifications at lower taxonomic ranks (Porter and Hajibabaei 2020).

Considering the importance of choice of marker and primer to improve taxonomic coverage and discrimination of DNA metabarcoding, we investigated the impact of these factors on the composition and structure of marine macrozoobenthic communities. We selected two different primer pairs targeting COI and one targeting 18S rDNA V4 region, to compare their ability to detect macroinvertebrates at species level and to evaluate the benefits of the use of two molecular markers on species recovery success. We also conducted an assessment of the availability of reference sequences for all species detected in the study, in order to compare the representativeness

of species, identifying the existence of gaps in both databases (BOLD for COI and SILVA for 18S rRNA gene) and attempting to infer the reasons for failed detections.

Materials and methods

Sampling design

This study was conducted in Ría de Vigo, a semi-enclosed heavily human populated bay on the NW coast of Spain, constituted by important busy harbours and consequently affected by several human activities (e.g. sewage runoff or harvesting) (Veiga et al. 2016). This area includes both hard and soft substrata, which have a high primary productivity due to the influence of coastal seasonal upwelling-downwelling dynamics, which lead the larvae recruitment process during spring and early summer (Prego and Fraga 1992).

In December 2016, four replicates (flat panels 10 × 10 cm) of three different types of artificial substrates – slate, polyvinyl chloride (PVC) and granite – were randomly deployed on the dock of Toralla Island (42°12'2.267"N, 8°48'4.187"W) approximately 1.5 m of depth (Suppl. material 1: Fig. S1). Using a sterile hermetic plastic bag, after 3, 7, 10 and 15 months, one replicate of each substrate was randomly removed. At the laboratory, the content of each bag with substrate plates was placed in a tray. Then, each sample was individually photographed and the representative mobile and sessile fauna were separated. Each sample was washed with filtered sea water, and the mobile fauna was sieved using a 500 µm mesh. After collecting the mobile fauna, the sessile fauna was scraped with a spatula into a tray. The water of each tray was also sieved (500 µm mesh) and preserved with mobile fauna samples. Between samples, all materials were properly sterilized. All samples were then preserved in ethanol and stored at -20 °C until further analysis.

DNA extraction, PCR amplification and HTS procedures

We extracted the DNA from the bulk biomass using DNA extraction procedures adapted from Ivanova et al. (2006) silica-based method, as described in Steinke et al. (2021). The mobile and sessile fauna were processed separately, including amplification and sequencing. Ethanol preserved samples were first filtered to retain the biomass and the ethanol was discarded. The filtered biomass was left for four hours in the hotte to allow the ethanol to fully evaporate. Then, based on the wet weight of each sample (as suggested by Steinke et al. 2021), the appropriate volume of lysis buffer solution (100 mM NaCl, 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.5% SDS) was added and the samples were incubated overnight at 56 °C gently mixed in an orbital shaker set. Negative controls were included in DNA extraction procedures and later checked together with the regular samples during DNA amplification. To maximize diversity recovery, two aliquots of each lysate were used, totaling two DNA extractions per sample. After extraction, the aliquots of genomic DNA for the same sample were pooled before amplification and HTS.

The production of amplicon libraries and the HTS were carried out at Genoinseq (Cantanhede, Portugal). A preliminary assessment of primer amplification efficiency of COI was conducted in a subset of samples to test four primer pairs that have been previously used in DNA metabarcoding studies (more details in Suppl. material 1: Table S1 and Figs S2, S3). Together, the primer pairs mlCOIintF/LoboR1 and LCO1490/Ill_C_R recovered the highest species richness corresponding to more than 85% of the total detected species with COI (data in Suppl. material 1). Therefore, these two primer pairs were selected, and combined with one primer pair targeting the 18S V4 rRNA gene to amplify the marine macroinvertebrate communities from each sample (Table 1). The 18S V4 primer (TAReuk454FWD1/TAReukREV3) was selected based on results obtained in other studies, in which different 18S primers were compared, and this primer had the best performance (Fais et al. 2020b).

PCR reactions were performed using KAPA HIFI HotStart PCR Kit for the COI primer pair without inosines (mlCOIintF/LoboR1) and for the 18S V4 region primer. PCR amplification reactions contained 0.3 µM of each primer and 50 ng of template DNA in the case of COI amplification and 12.5 ng for 18S V4 amplification, in a total volume of 25 µL. For the other COI primer pair (LCO1490/Ill_C_R), PCR reactions were performed using 1× Advantage 2 Polymerase Mix (Clontech, Mountain View, CA, USA), 0.2 µM of each PCR primer and 25 ng of DNA template in a total volume of 25 µL. Second PCR reactions added indexes and sequencing adapters to both ends of the amplified target region (MiSeq Reagent Kit v3 – 600-cycle) according to manufacturer's recommendations (Illumina 2013). PCR products were then one-step purified and normalized using SequalPrep

Table 1. Primer pairs and respective thermal cycling conditions used in this study to amplify marine macroinvertebrate communities. F – forward; R – reverse; bp – base pairs.

Primer combinations and length		Direction (5'-3')	Reference	PCR thermal cycling conditions
COI	LCO1490/ Ill_C_R (325 bp)	(F) GGTCAACAAATCATAAAGATATTGG	Folmer et al. 1994	(1) 94 °C (5 min); (2) 35 cycles: 94 °C (30 s), 52 °C (90 s), 68 °C (60 s); (3) 68 °C (10 min).
		(R) GGIGGRTAIACIGTTCAICC	Shokralla et al. 2015	
	mlCOIintF/LoboR1 (313 bp)	(F) GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al. 2013	(1) 95 °C (3 min); (2) 35 cycles: 98 °C (20 s), 60 °C (30 s), 72 °C (30 s); (3) 72 °C (5 min).
		(R) TAAACYTCWGGRTGWCCRAARAAYCA	Lobo et al. 2013	
18S	TAReuk454FWD1/ TAReukREV3 (400 bp)	(F) CCAGCASCYGCGGTAATTCC	Stoeck et al. 2010; Lejzerowicz et al. 2015	(1) 95 °C (3 min); (2) 10 cycles: 98 °C (20 s), 57 °C (30 s), 72 °C (30 s); (3) 25 cycles: 98 °C (20 s), 47 °C (30 s), 72 °C (30 s); (4) 72 °C (5 min).
		(R) ACTTTCGTTCTTGATYRA		

96-well plate kit (ThermoFisher Scientific, Waltham, USA) (Comeau et al. 2017), pooled and paired-end (2× 300 bp) sequenced in an Illumina MiSeq sequencer with the V3 chemistry, according to manufacturer's instructions (Illumina, San Diego, CA, USA).

Negative and positive controls were included in PCR amplification. As positive controls, we used a DNA extract previously tested successfully for PCR. Success of PCR amplification was checked by electrophoresis. No amplification was detected in any of the negative controls from DNA extraction or PCR.

Amplification failed with the primer mlCOIintF/LoboR1 in the sample of mobile fauna of the granite substrate after 3 months of deployment and, consequently, was not considered for further analysis.

Data processing

Raw reads in fastq format generated by MiSeq sequencing were quality-filtered with PRINSEQ version 0.20.4 (Schmieder and Edwards 2011). Sequencing adapters and reads with less than 100 bp (for the COI region) and with less than 150 bp (for the 18S V4 region) were removed. Bases with an average quality lower than Q25 in a sliding window of 5 bases were trimmed. Then, the filtered forward and reverse reads obtained were merged (make.contigs function, default alignment) by overlapping paired-end reads in mothur 1.39.5 (Schloss et al. 2009; Kozich et al. 2013), and primers were removed (trim.seqs function, default).

The usable reads were then processed in two pipelines of public databases: a) COI reads were submitted to mBrave – Multiplex Barcode Research and Visualization Environment (www.mbrave.net; Ratnasingham 2019), using the sample batch function which is linked with BOLD (Ratnasingham and Hebert 2007); b) 18S reads were analyzed in SILVAngs database (https://ngs.arb-silva.de/silvangs/; Quast et al. 2013).

In mBrave, since primer sequences were previously removed in mothur, only the trimming by length was applied (maximum 313 bp for mlCOIintF/LoboR1 and 325 bp for LCO1490/Ill_C_R; minimum 150 bp) and only reads with minimum quality value (QV) higher than 10 were kept. This filtering step allowed for a max of 25% nucleotides with <20 QV value and max 25% nucleotides with <10 QV value. Reads were then taxonomically assigned at species level using a 97% similarity threshold against BOLD database that includes several publicly available reference libraries for marine invertebrates of the northeast Atlantic (e.g. Hollatz et al. 2017; Leite et al. 2020; Vieira et al. 2021).

Output fasta files produced in mothur for the 18S marker were then processed by the amplicon analysis pipeline of the SILVA project (SILVAngs 1.4; Quast et al. 2013). Each read was aligned using the SILVA Incremental Aligner (SINA v1.2.10 for ARB SVN (revision 21008); Pruesse et al. 2012) against the SILVA SSU rRNA SEED and quality controlled (Quast et al. 2013). Reads shorter than 150 aligned nucleotides and reads with more than 2% of ambiguities or homopolymers were excluded from further processing. Putative contaminations and artefacts, reads with a low alignment quality (50 alignment identity, 40 alignment score reported by SINA), were identified and excluded from downstream analysis. After these initial steps, identical reads were identified (dereplicated), the unique reads were clustered (OTUs), on a per sample basis, and the reference read of each OTU was classified. Dereplication and clustering was done using VSEARCH (version 2.15.1; https://github.com/torognes/vsearch; Rognes et al. 2016) applying identity criteria of 1.00 and 0.99, respectively. The classification was performed by a local nucleotide BLAST search against the non-redundant version of the SILVA SSU Ref dataset (release 138.1; http://www.arb-silva.de) using blastn (version 2.2.30+; http://blast.ncbi.nlm.nih.gov/Blast.cgi) with standard settings (Camacho et al. 2009). The classification of each OTU reference read was mapped onto all reads that were assigned to the respective OTU using a 99% similarity threshold. Reads without any classification or reads with weak BLAST hits, where the function "(% sequence identity + % alignment coverage)/2" did not exceed the value of 99, remain unclassified. These reads were assigned to "No Taxonomic Match".

For both markers, only reads with match at species level were used for further analysis, and taxonomic assignments with less than 8 sequences were discarded. Any read that matched to non-metazoan was also excluded. The validity of the species names was verified in the World Register of Marine Species (WoRMS) database (WoRMS Editorial Board 2021). The obtained reads were analyzed separately for mobile and sessile fauna samples, and then combined for data analysis.

Revision of the species' matches accuracy and comparison of the representativeness of the detected species in the two databases

Incongruences in genetic databases are an ongoing problem that can affect taxonomic assignments (Hleap et al. 2021). Although some efforts have been developed to solve this issue (e.g. Fontes et al. 2021; Radulovici et al. 2021), incongruences still persist in genetic databases. To have a broad-range of representativeness and maximize results, the taxonomic assignment was made using the full genetic databases (i.e. BOLD and SILVA) and not a curated in-house reference library. However, the confidence and accuracy in the taxonomic assignments can be compromised. For that reason, we reviewed individually each species match to assess the reliability of the taxonomic assignments. Discordances were carefully inspected and if they were possible to resolve (e.g. synonyms, clear cases of misidentification), the most probable identification was kept.

We then assessed the presence of representative sequences of all the species detected in the present study in BOLD and SILVA. Failed detection by one marker may simply have occurred because that particular species was

not present in the respective reference database. However, if a species was present in both databases, but was only detected by one marker, that would be an indication of probable PCR amplification failure of the marker that failed detection. All the available COI sequences matching the detected species names were mined from BOLD using BAGS (Fontes et al. 2021). To assess which species have representative sequences in SILVA, all the Animalia records were mined directly from the database (version 138.1). A species was considered represented if at least one sequence was available.

Molecular markers comparison

The proportion of species with overlapping or exclusive detections by each primer pair and marker was determined for all substrates and sampling time combinations, using Venn diagrams (http://www.venndiagrams.net/). For each primer pair the distribution of species among high-rank taxonomic groups (e.g. order or phyla) was displayed through barplots (GraphPad Software, Inc.).

To identify clusters of data objects (species level identifications) in the dataset, the unsupervised machine learning k-means was applied, in the presence/absence matrix of the global species detected by primer pairs and markers, which groups the data without prior categories. The optimal number of clusters was determined with the elbow (fviz_nbclust, method = "wss" function) and silhouette (silhouette function) analysis. The analyses were performed (kmeans function) and visualized (fviz_cluster function) in R with the packages "cluster" (Maechler et al. 2019), "factoextra" (Kassambara and Mundt 2020) and "stats" (R core team 2020) as default.

Results

Effect of marker and primer choice on species detection

High-throughput sequencing of marine macroinvertebrate samples, for both markers and three primer pairs, generated a total of 2,956,328 raw reads. Following bioinformatic processing, a total of 2,356,818 reads were retained (Table 2). Of these, 55.6% were assigned to a marine macroinvertebrate species: 46.8% using mlCOIintF/LoboR1, 36.6% with LCO1490/Ill_C_R and 16.6% with TAReuk454FWD1/TAReukREV3 (Suppl. material 2: Table S2). Of the remaining reads, 0.1% were rare OTUs (<8 reads) and 44.3% could not be assigned to macrozoobenthic species or to a metazoan phylum.

From the three types of artificial substrates sampled at four different deployment periods (12 samples), the three primer pairs were able to identify a total of 161 species, distributed by 9 taxonomic groups: Annelida, Bryozoa, Crustacea, Echinodermata, Hydrozoa, Mollusca, Nemertea, Platyhelminthes and Tunicata (species names and the associated taxonomic group displayed in Suppl. material 2: Table S3). The number of species detected was similar between primer pairs, with the primer pair TAReuk454FWD1/TAReukREV3 (18S V4 region) retrieving a total of 77 species, whereas among the COI primer pairs, mlCOIintF/LoboR1 allowed the detection of more species than LCO1490/Ill_C_R (84 species and 63 species, respectively).

Table 2. Total number of sequences generated in Illumina MiSeq high-throughput sequencing (raw reads), retained along processing steps of the bioinformatics pipeline (primers removal, demultiplex and quality filter), and assigned to taxonomic groups for each primer pair (mlCOIintF/LoboR1; LCO1490/Ill_C_R; TAReuk454FWD1/TAReukREV3).

	Primer pairs					
	mlCOIintF/ LoboR1		LCO1490/ Ill_C_R		TAReuk454FWD1/ TAReukREV3	
Raw reads	1110851	100,00%	945639	100,00%	899838	100,00%
First quality-filter*	953733	85,86%	808234	85,47%	594851	66,11%
After filtering**	953704	85,85%	798645	84,46%	581220	64,59%
Usable sequences***	869015	78,23%	587794	62,16%	411782	45,76%
Metazoa	655097	68,69%	579857	61,32%	218416	24,27%
No taxonomic match****	41641	3,75%	99367	10,51%	193366	21,49%
<8 sequences*****	287	0,03%	281	0,03%	986	0,11%
Species level taxonomic assignment	613169	55,20%	480209	50,78%	217430	24,16%

* primers removal, demultiplex and quality filter.
** Filtered reads: rejected sequences based on length, presence of ambiguities and homopolymers, putative contaminations and artifacts.
*** Reads submitted for taxonomic assignment.
**** Reads without taxonomic classification at species level at ≥97% for COI primers and ≥99% for 18S V4.
*****Matches at species level, however, with <8 sequences per identification.

The applied primers also differed in their efficiency to recover particular taxonomic groups (Fig. 1). In the case of COI primer pairs, Crustacea, Mollusca and Annelida were the taxa with higher species diversity (77.3% for mlCOIintF/LoboR1 and 79.7% for LCO1490/Ill_C_R), while for TAReuk454FWD1/TAReukREV3 the most represented taxonomic groups were Annelida, Mollusca, Bryozoa and Hydrozoa (79% of the total detected taxa). Furthermore, whereas Tunicata and Platyhelminthes were only detected by 18S, Mollusca and Crustacea had more species identified with COI primers.

A higher species richness was detected consistently at seven months for all primer pairs in the three substrates. The species detected by each primer pair, and also the taxonomic groups, were different between primers and markers in the four sampling times and between substrates (Suppl. material 1: Fig. S4).

If the combined number of detected species by the two COI primer pairs is used, the 18S V4 region retrieved less taxa than COI (77 species vs 107 species, respectively). Both elbow and silhouette analysis retrieved two as the optimal number of k (i.e. two clusters; Suppl. material 1: Fig. S5). Two well segregated groups (Fig. 2) with no overlap were retrieved, one corresponding to the data belonging to COI and other to 18S detected species.

The two molecular markers and the three primer pairs used were highly complementary in their ability to detect marine macroinvertebrate species (Fig. 3). Among the detected species, only 8.1% were common

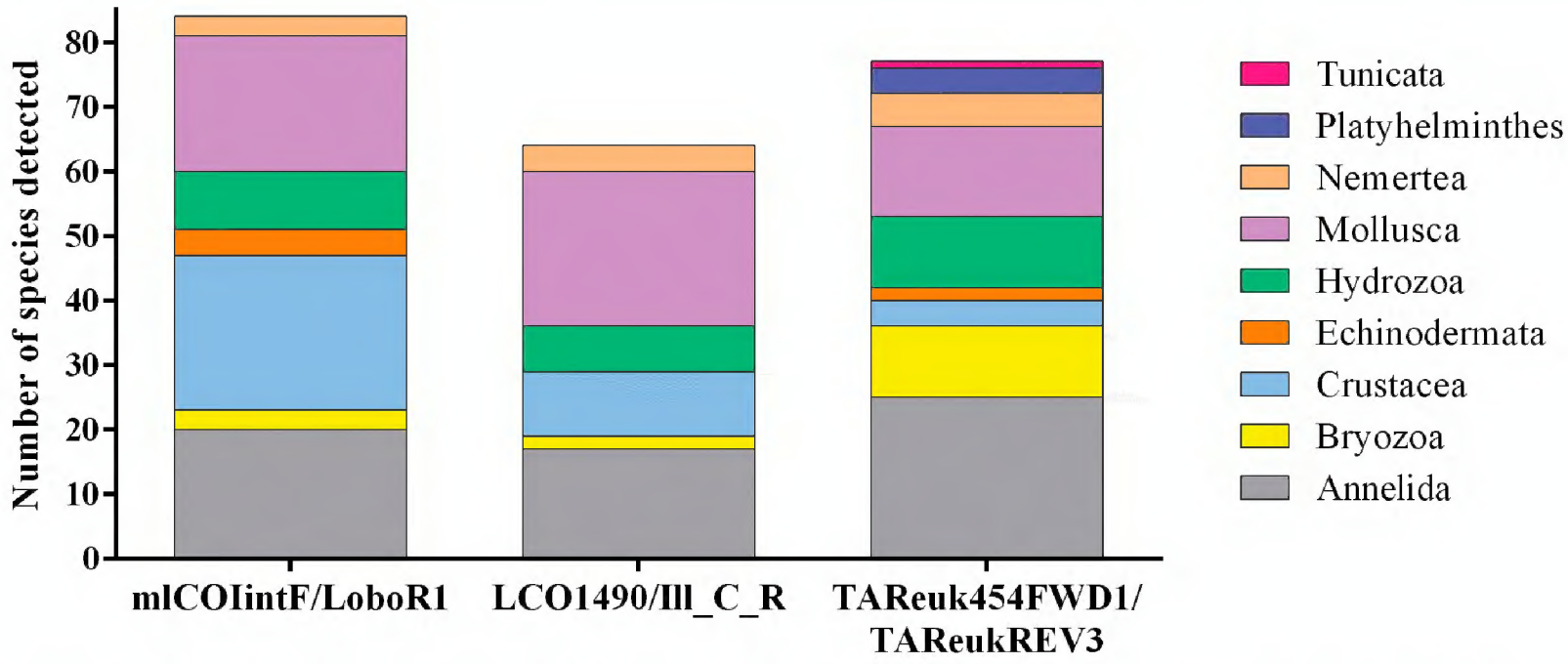

Figure 1. Taxonomic profile of the marine macroinvertebrate species detected in the substrates by each primer pair.

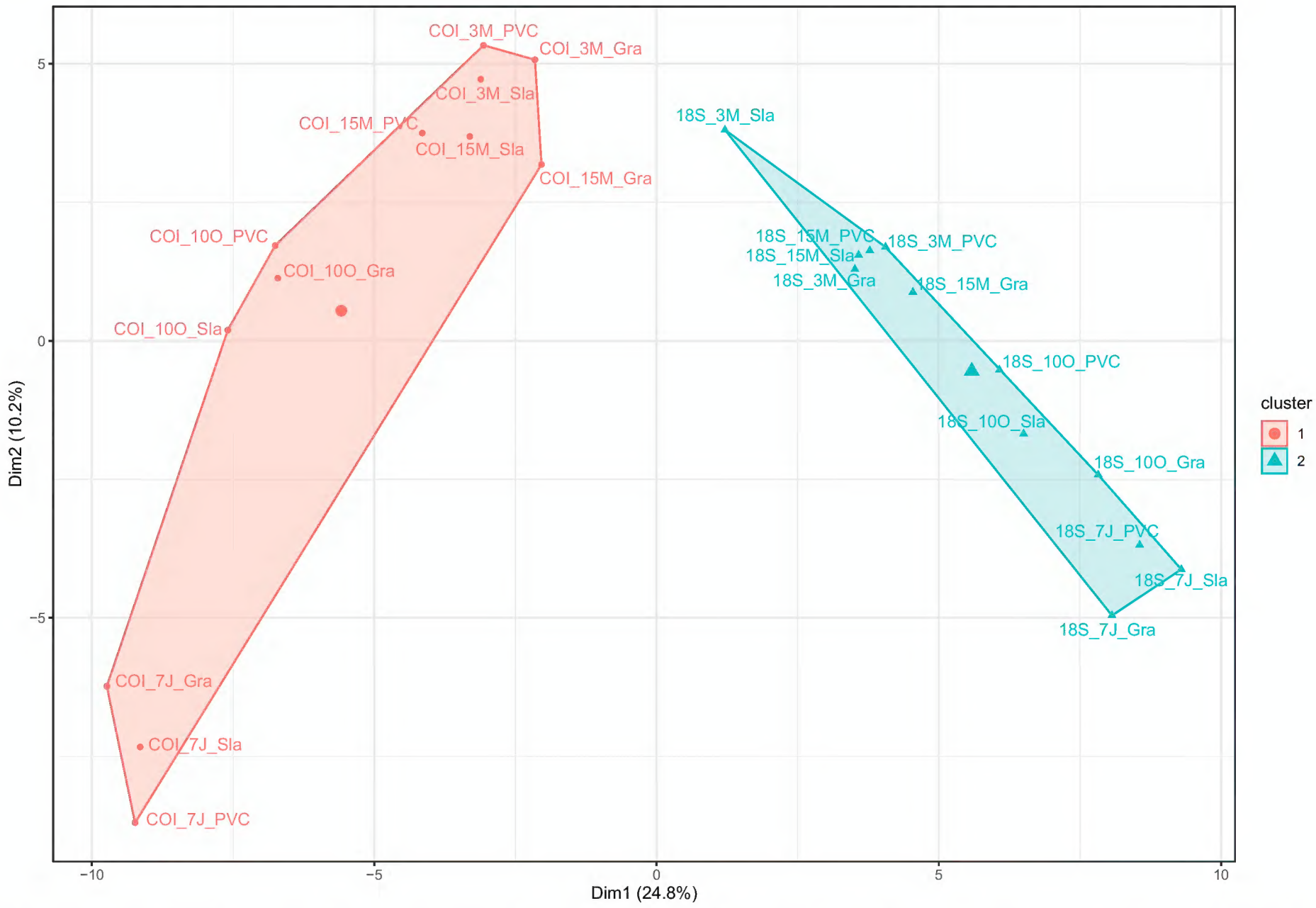

Figure 2. Best fitting number of clusters (k=2) using the unsupervised machine learning k-means for the combined identifications of marine macrozoobenthic species using both COI-primers (COI – red) and the 18S-primer (18S – blue), in the three sampling times (3M – 3 months; 7M – 7 months; 10M – 10 months; 15M – 15 months) and the three artificial substrates (Sla – slate; PVC; Gra – granite).

to the three primer pairs and 68.9% were exclusively recovered by one primer: 21.7% for mlCOIinF/LoboR1, 13.7% for LCO1490/Ill_C_R and 33.5% for TAReuk-454FWD1/TAReukREV3. For COI-primers, Crustacea was the taxon with more species exclusively detected by mlCOIintF/LoboR1 (45.7% of the species), whereas for LCO1490/Ill_C_R was Mollusca (50% of the species). For the 18S primer pair, 65.5% of the exclusively detected species were Annelida (31.2%), Mollusca (14.8%) and Bryozoa (18.5%).

Availability of reference sequences on public databases

Considering a total of 161 marine macroinvertebrate species detected combining together the results of 18S V4 and the two COI primers, we evaluated the taxonomic coverage in the respective databases, namely mBrave for COI and SILVA for 18S V4. As much as 18% of the species still lack representative sequences of COI and 28% of the V4 region of the 18S rRNA gene (Fig. 4). While Crustacea was the taxonomic group with the higher num-

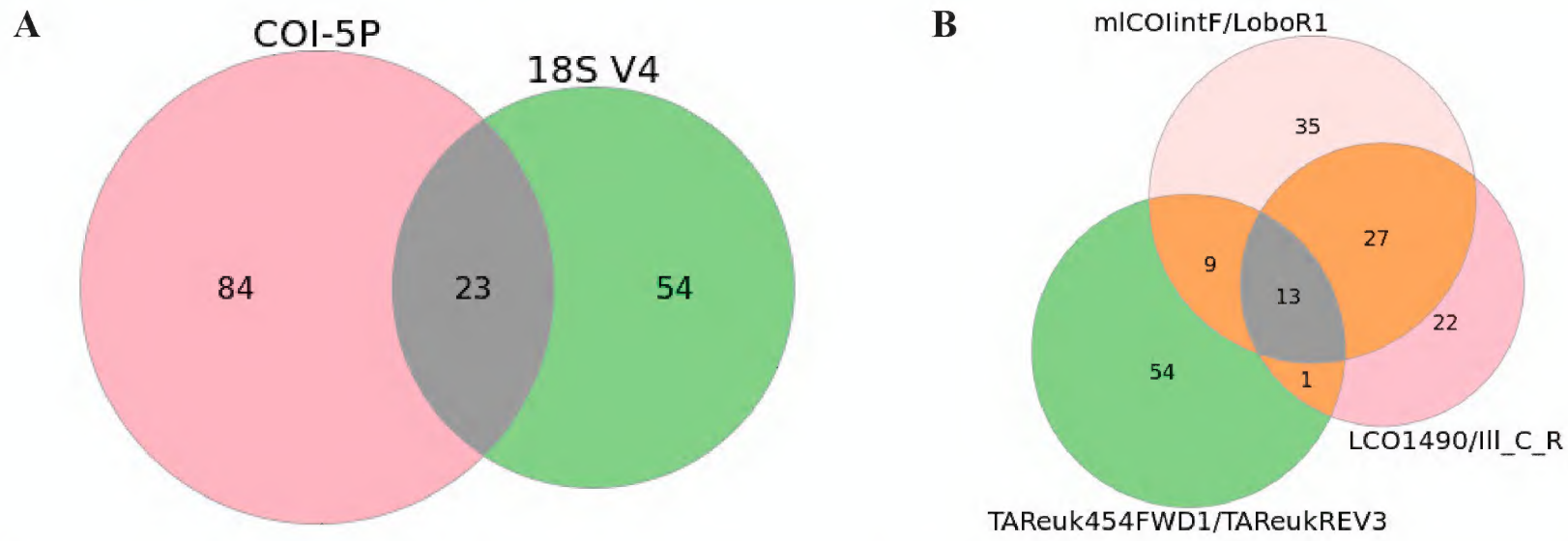

Figure 3. Partitioning of the marine macroinvertebrate species detection for (A) both marker loci and (B) primer pair, in the three substrate types and among all sampling times.

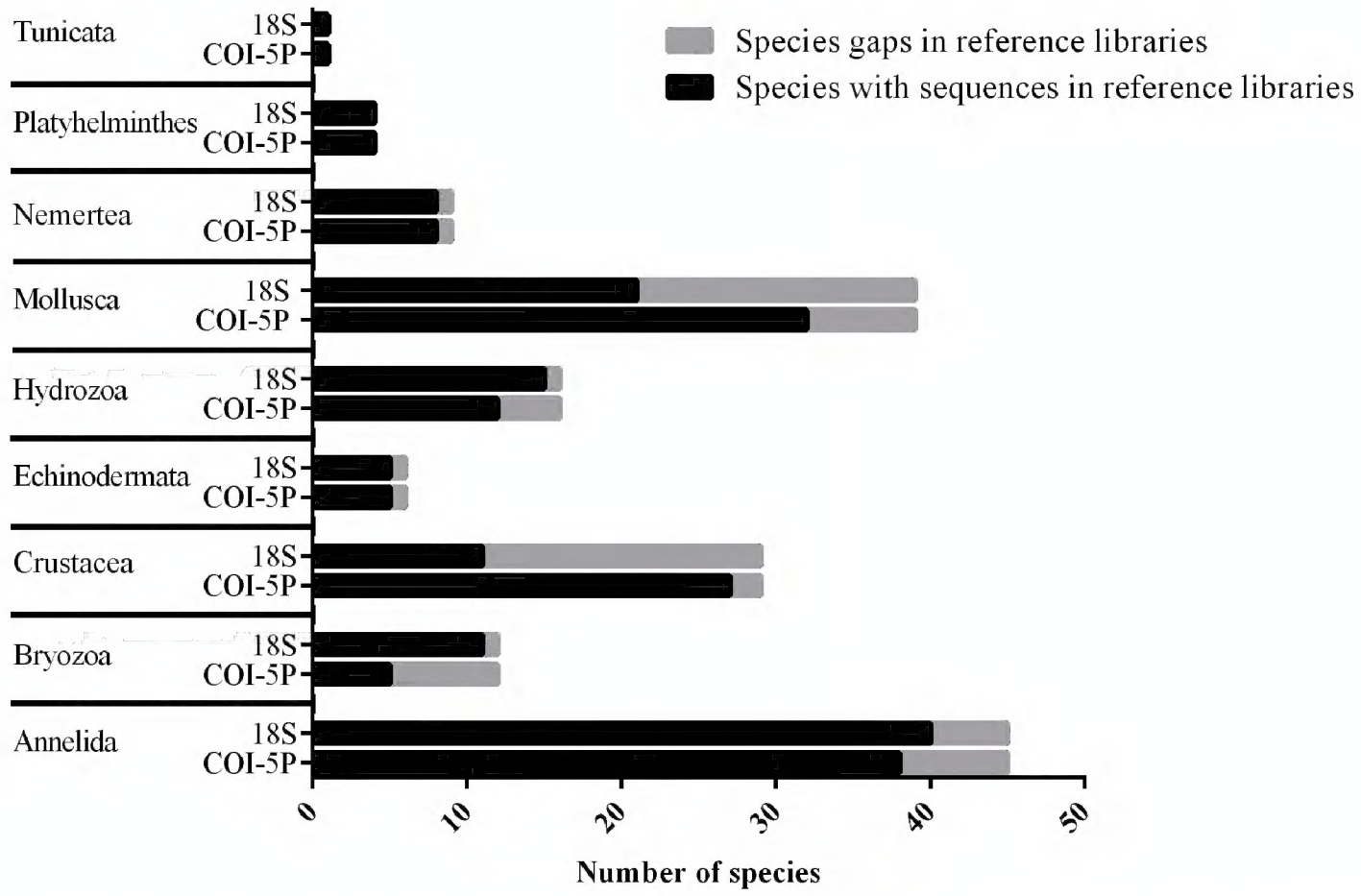

Figure 4. Availability of reference sequences of COI and 18S V4 for each taxonomic group of marine macroinvertebrate species detected with the three primer pairs from COI and 18S genes. Barcode coverage with at least one sequence per species (black bar).

ber of missing sequences in SILVA (38% of the total crustacean species detected), in BOLD, Bryozoa was the group with more species missing COI sequences (42% of the total bryozoans species detected).

Since taxonomic assignment can be affected by incongruences in genetic databases, a manual inspection of the taxonomic assignments may be advisable to more accurately compare results. Overall, the fair majority of the species assignments (94%) appear to have a high level of certainty (Suppl. material 2: Table S3). There were only a minority of cases where species could not be confidently discriminated with currently available data (e.g. *Turritopsis nutricula/T. dohrnii*).

Discussion

DNA metabarcoding-based biomonitoring of aquatic communities would benefit from the establishment of standardized approaches (Blackman et al. 2019; Duarte et al. 2021; Gielings et al. 2021), which would enable direct inter-study comparisons, broader implementation and greater scientific gains. Choice of marker and primer pairs to employ is a critical decision in DNA metabarcoding, that depends on the goals of the study, targeted communities and aimed rank of taxonomic assignment (van der Loos and Nijland 2020; Duarte et al. 2021). Although criteria of choice such as practical efficiency and cost-effectiveness would favor finding the "best" single primer pair that fits a particular study, our results indicated that, when the goal is to assess as comprehensively as possible the species composition of a marine macrozoobenthic sample, the ideal primer pair may be hard to find. Both biased primer affinity for the DNA templates and gaps in species representativeness in reference databases appear to concur for this difficulty.

Together, the three primer pairs used in this study enabled the detection of a fair number of macrozoobenthic

species. Quantitatively, the 18S V4 region and the two COI primer pairs displayed similar ability to detect marine macroinvertebrate species. However, considering the high complementarity in the species recovered between these markers, the choice of the best performing primer was not obvious. Similar COI – 18S comparisons described in the literature report somewhat distinct and even contradictory results. For example, DNA metabarcoding studies using mock zooplankton communities demonstrated different taxonomic recovery ability between COI and 18S: whereas one of the studies detected similar patterns of species detection ability among markers (Clarke et al. 2017), the other reported higher detection efficiency with 18S V4 region than with COI, and with higher levels of overlapping between markers (Zhang et al. 2018). Contrary to these results, in a study using marine hard-bottom communities the authors showed that using COI (Leray-XT) they were able to recover four times more diversity than using 18S region (V7) (Wangensteen et al. 2018). However, a direct comparison with our study is limited given the employment of different primers, reference libraries and thresholds. The differences obtained for the 18S V4 region in assignment for some taxonomic groups could be related to more conserved sequences in priming regions (Tang et al. 2012; Brown et al. 2015). In addition, taxonomic discrimination efficiency of the targeted marker loci, possible mismatches between primer and DNA template or difficulties in primer affinity for specific taxa can be possible reasons for the obtained differences between primers and marker region in amplification efficiency.

The significant complementarity observed between the two molecular markers, with each single marker capturing at the very best approximately 66% of the detected species in a sample, revealed that, by using only one marker, a fair portion of the marine macroinvertebrate species may fail detection. Both markers detected different communities, raising high concerns for monitoring studies, since the biodiversity detected will be different and many species may be overlooked due to methodological steps only. For example, while isopods were only detected by COI, platyhelminthes and tunicates were exclusively detected by 18S V4. Although few metabarcoding studies compared the performance of molecular markers on species recovery (Dowle et al. 2015; Drummond et al. 2015), our results are generally consistent with previous findings in marine invertebrate communities (Wangensteen et al. 2018), where high levels of complementarity at species level between these markers were reported. Since our goal is to capture the widest possible diversity of macrozoobenthic species, compared to using a single molecular marker, the combination of a multi-locus strategy improves the number of retrieved species, which we recommend as the best practice to be used in marine macroinvertebrates assessments.

Regarding the two primer pairs from the COI barcode region, although we observed a similar number of species globally detected by each primer (84 vs 63 species), they diverge qualitatively in the species detected (41% of the species exclusively detected with mlCOIintF/LoboR and 21% in LCO1490/Ill_C_R), hence this should be the main criteria to consider in order to maximize the scope of species detection. The efficiency of different COI-primers in macroinvertebrates assessment has been already compared in previous studies (Hollatz et al. 2017; Lobo et al. 2017; Ip et al. 2019; Derycke et al. 2020), however not always reaching the same conclusions. Different methodologies adopted in different studies may explain the discrepancies in the results (e.g.: PCR thermal cycling regimes, sequencing depths, informatics platforms, clustering and threshold values, reference libraries). Our results emphasize that using a single primer pair for the COI region will result in a fair amount of undetected diversity of marine zoobenthic taxa. Therefore, we suggest the simultaneous employment of at least two primer pairs to improve the efficiency of the taxonomic diversity recovery.

The three primer pairs used in this study were able to detect marine macroinvertebrate species in every sampling time, all of them consistently pointing to a higher species diversity after seven months of deployment of the substrates. These results highlight the benefit of the application of a multiple primer pair and multi-locus strategy for ecological assessments of marine species, since if we had only used one primer pair or marker we would have failed to detect important macrobenthic taxa, and the taxonomic composition of the community could emerge substantially different. Temporal and seasonal changes in a community could affect the potential of DNA-based species monitoring, especially when methodological bias originated by amplification procedures (choice of marker loci and primer pairs) could influence ecological interpretations (Clarke et al. 2017).

We performed an assessment of the availability of representative sequences for all species detected in this study in each of the reference databases employed, namely BOLD and SILVA, respectively for COI and 18S V4 markers. This enabled us to verify if the detection of a species by one marker, and not by the other, could be attributed to gaps in the library of the latter or, if no gap was found, it could be ascribed to faulty amplification. A sizable but minor proportion of gaps was recorded for both markers (18% for COI and 28% for 18S V4). The incompleteness, and possible inaccuracies of databases may explain some of the species detected exclusively by one marker, as for example, the flatworm *Vorticeros auriculatum*, a species detected by 18S V4 that does not have representatives in BOLD. On other hand, some of the detected species with reference sequences in both databases were only detected by one marker (e.g. the tunicate *Asterocarpa humilis*, undetected with COI despite having representative sequences in BOLD). Considering the complete 18S rRNA gene, the target region we selected should not be the main reason for failed detections, since V4 is reported to have high amplification success (Brown et al. 2015; Lejzerowicz et al. 2015; Zhang et al. 2018) and demonstrated to have a better performance on taxonomic assignments than other 18S regions (Fais et al. 2020b). However, since these primers failed to detect some species, this was probably due to faulty amplification or reduced resolution of taxonomic assignments. Indeed, although COI-based monitoring approaches may claim the advantage of having a verified and dedicated database

for a large variety of taxa (Porter and Hajibabaei 2020), several studies already reported the existence of significant gaps in reference libraries particularly for dominant marine macrobenthic taxa (Weigand et al. 2019: 20% to 30% of completion of databases for dominant groups, i.e. Annelida, Mollusca and Arthropoda), including for the region that comprised the geographic area of the current study (Leite et al. 2020: 49% for Mollusca, 53% for Crustacea and 16% for Polychaeta). Hence, although the availability of sequences or reference libraries does not appear to have been the main factor affecting species detection, it revealed that more investment should be allocated to obtain reliable reference sequences to enhance species assignment accuracy, in order to recover the taxonomic composition of a target community as complete as possible.

Globally, these results highlight the influence of marker and primer pair complementarity on the ability to record marine macrozoobenthic species through metabarcoding. For future high-throughput assessments using DNA metabarcoding approaches, we recommend combining molecular markers and, if possible, multiple primer pairs, to increase the power of species detections and the accuracy of biodiversity assessments, thereby yielding more comprehensive and reliable results for marine macroinvertebrate monitoring.

Author contributions

Conceptualization: B.R.L., J.S.T. and F.O.C.; Methodology: B.R.L and P.E.V.; Formal analysis: B.R.L. and P.E.V.; Data curation: B.R.L and P.E.V.; Writing – original draft: B.R.L. and F.O.C.; Writing – review and editing: B.R.L., P.E.V., J.S.T. and F.O.C.; Visualization: B.R.L. and P.E.V. Supervision: J.S.T. and F.O.C. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

This work was supported by the project ATLANTIDA – Platform for the monitoring of the North Atlantic Ocean and tools for the sustainable exploitation of the marine resources, with the reference NORTE-01-0145-FEDER-000040, co-financed by the European Regional Development Fund (ERDF), through Programa Operacional Regional do Norte (NORTE 2020). BRL benefitted from an FCT fellowship PD/BD/127994/2016. The authors would like to thank Sofia Duarte (University of Minho) for the availability and support during practical stages of the research.

References

Andújar C, Arribas P, Yu DW, Vogler AP, Emerson BC (2018) Why the COI barcode should be the community DNA metabarcode for the metazoa. Molecular Ecology 27: 3968–3975. https://doi.org/10.1111/mec.14844

Atienza S, Guardiola M, Præbel K, Antich A, Turon X, Wangensteen OS (2020) DNA metabarcoding of deep-sea sediment communities using COI: community assessment, spatio-temporal patterns and comparison with 18S rDNA. Diversity 12(4): 123. https://doi.org/10.3390/d12040123

Aylagas E, Borja Á, Muxika I, Rodríguez-Ezpeleta N (2018) Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. Ecological Indicators 95: 194–202. https://doi.org/10.1016/j.ecolind.2018.07.044

Bista I, Carvalho GR, Tang M, Walsh K, Zhou X, Hajibabaei M, Shokralla S, Seymour M, Bradley D, Liu S, Christmas M, Creer S (2018) Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. Molecular Ecology Resources 18: 1020–1034. https://doi.org/10.1111/1755-0998.12888

Blackman RC, Mächler E, Altermatt F, Arnold A, Beja P, Boets P, Egeter B, Elbrecht V, Filipe AF, Jones JI, Macher J, Majaneva M, Martins FMS, Murria C, Meissner K, Pawlowski J, Schmidt Y, Paul L, Zizka VMA, Leese F, Price BW, Deiner K (2019) Advancing the use of molecular methods for routine freshwater macroinvertebrate biomonitoring–the need for calibration experiments. Metabarcoding and Metagenomics 3: e34735. https://doi.org/10.3897/mbmg.3.34735

Brown EA, Chain FJJ, Crease TJ, MacIsaac HJ, Cristescu ME (2015) Divergence thresholds and divergent biodiversity estimates: Can metabarcoding reliably describe zooplankton communities? Ecology and Evolution 5: 2234–2251. https://doi.org/10.1002/ece3.1485

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T (2009) BLAST+: architecture and applications. BMC Bioinformatics 10: e421. https://doi.org/10.1186/1471-2105-10-421

Castro LR, Meyer RS, Shapiro B, Shirazi S, Cutler S, Lagos AM, Quirosa SY (2021) Metabarcoding meiofauna biodiversity assessment in four beaches of Northern Colombia: effects of sampling protocols and primer choice. Hydrobiologia 848: 3407–3426. https://doi.org/10.1007/s10750-021-04576-z

Chang JJM, Ip YCA, Bauman AG, Huang D (2020) MinION-in-ARMS: Nanopore sequencing to expedite barcoding of specimen-rich macrofaunal samples from Autonomous Reef Monitoring Structures. Frontiers in Marine Science 7: e448. https://doi.org/10.3389/fmars.2020.00448

Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. Ecology and Evolution 7(3): 873–883. https://doi.org/10.1002/ece3.2667

Clarke LJ, Soubrier J, Weyrich LS, Cooper A (2014) Environmental metabarcodes for insects: *in silico* PCR reveals potential for taxonomic bias. Molecular Ecology Resources 14: 1160–1170. https://doi.org/10.1111/1755-0998.12265

Collins RA, Bakker J, Wangensteen OS, Soto AZ, Corrigan L, Sims DW, Genner MJ, Mariani S (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution 10(11): 1985–2001. https://doi.org/10.1111/2041-210X.13276

Comeau AM, Douglas GM, Langille MGI (2017) Microbiome Helper: a custom and streamlined workflow for microbiome research. mSystems 2: e00127-16. https://doi.org/10.1128/mSystems.00127-16

Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Miné J, Arnaud-Haond S (2015) Metabarcoding is powerful yet still blind: A comparative analysis of morphological and molecular surveys of seagrass communities. PLoS ONE 10(2): e0117562. https://doi.org/10.1371/journal.pone.0117562

Derycke S, Maes S, van den Bulcke L, Vanhollebeke J, Wittoeck J, Hillewaert H, Ampe B, Haegeman A, Hostens K, de Backer A (2020) Optimization of metabarcoding for monitoring marine macrobenthos: primer choice and morphological traits determine species detection in bulk DNA and eDNA from the ethanol preservative. Authorea. https://doi.org/10.22541/au.159665093.39993653

Dowle EJ, Pochon X, Banks J, Shearer K, Wood SA (2015) Targeted gene enrichment and high throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. Molecular Ecology Resources 16(5): 1240–1254. https://doi.org/10.1111/1755-0998.12488

Drummond AJ, Newcomb RD, Buckley TR, Xie D, Dopheide A, Potter BCM, Heled J, Ross HA, Tooman L, Grosser S, Park D, Demetras NJ, Stevens MI, Russell JC, Anderson SH, Carter A, Nelson N (2015) Evaluating a multigene environmental DNA approach for biodiversity assessment. Gigascience 4(1): s13742-015. https://doi.org/10.1186/s13742-015-0086-1

Duarte S, Leite BR, Feio MJ, Costa FO, Filipe AF (2021) Integration of DNA-based approaches in aquatic ecological assessment using benthic macroinvertebrates. Water 13(3): 331. https://doi.org/10.3390/w13030331

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Hebert PDN, Steinke D (2019) Validation of COI metabarcoding primers for terrestrial arthropods. PeerJ 7: e7745. https://doi.org/10.7717/peerj.7745

Elbrecht V, Leese F (2017) Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science 5: 1–11. https://doi.org/10.3389/fenvs.2017.00011

Fais M, Bellisario B, Duarte S, Vieira PE, Sousa R, Canchaya C, Costa FO (2020a) Meiofauna metabarcoding in Lima estuary (Portugal) suggests high taxon replacement within a background of network stability. Regional Studies in Marine Science 38: 101341. https://doi.org/10.1016/j.rsma.2020.101341

Fais M, Duarte S, Vieira PE, Sousa R, Hajibabaei M, Canchaya C, Costa FO (2020b) Small-scale spatial variation of meiofaunal communities in Lima estuary (NW Portugal) assessed through metabarcoding. Estuarine, Coastal and Shelf Science 238: 106683. https://doi.org/10.1016/j.ecss.2020.106683

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3(5): 294–299.

Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: an automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21: 573–583. https://doi.org/10.1111/1755-0998.13262

Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources 13: 851–861. https://doi.org/10.1111/1755-0998.12138

Giebner H, Langen K, Bourlat SJ, Kukowka S, Mayer C, Astrin JJ Misof B, Fonseca VG (2020) Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. Molecular Ecology Resources 20(5): 1333–1345. https://doi.org/10.1111/1755-0998.13201

Gielings R, Fais M, Fontaneto D, Creer S, Costa FO, Renema W, Macher JN (2021) DNA metabarcoding methods for the study of marine benthic meiofauna: a review. Frontiers in Marine Sciences 8: 730063. https://doi.org/10.3389/fmars.2021.730063

Haenel Q, Holovachov O, Jondelius U, Sundberg P, Bourlat S (2017) NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hållö island, Smögen, and soft mud from Gullmarn Fjord, Sweden. Biodiversity Data Journal 5: e12731. https://doi.org/10.3897/BDJ.5.e12731

Hajibabaei M, Porter TM, Wright M, Rudar J (2019) COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. PLoS ONE 14: e0220953. https://doi.org/10.1371/journal.pone.0220953

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences 270(1512): 313–321. https://doi.org/10.1098/rspb.2002.2218

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21: 2190–2203. https://doi.org/10.1111/1755-0998.13407

Hollatz C, Leite BR, Lobo J, Froufe H, Egas C, Costa FO (2017) Priming of a DNA metabarcoding approach for species identification and inventory in marine macrobenthic communities. Genome 60(3): 260–271. https://doi.org/10.1139/gen-2015-0220

Illumina (2013) 16S metagenomic sequencing library preparation guide. https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf

Ip YCA, Tay YC, Gan SX, Ang HP, Tun K, Chou LM, Huang D, Meier R (2019) From marine park to future genomic observatory? Enhancing marine biodiversity assessments using a biocode approach. Biodiversity Data Journal 7: e46833. https://doi.org/10.3897/BDJ.7.e46833

Ivanova NV, deWaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Molecular Ecology Notes 6: 998–1002. https://doi.org/10.1111/j.1471-8286.2006.01428.x

Jungbluth MJ, Burns J, Grimaldo L, Slaughter A, Katla A, Kimmerer W (2021) Feeding habits and novel prey of larval fishes in the northern San Francisco Estuary. Environmental DNA 3: 1059–1080. https://doi.org/10.1002/edn3.226

Kassambara A, Mundt F (2020) Factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Kebschull JM, Zador AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Research 43(21): e143. https://doi.org/10.1093/nar/gkv717

Knittel T, Picard D (1993) PCR with degenerate primers containing deoxyinosine fails with Pfu DNA polymerase. Genome Research 2(4): 346–347. https://doi.org/10.1101/gr.2.4.346

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology 79: 5112–5120. https://doi.org/10.1128/AEM.01043-13

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution 33(7): 1870–1874. https://doi.org/10.1093/molbev/msw054

Leite BR, Vieira PE, Teixeira MAL, Lobo-Arteaga J, Hollatz C, Borges LMS, Duarte S, Troncoso JS, Costa FO (2020) Gap-analysis and

annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. Regional Studies in Marine Science 36: 101307. https://doi.org/10.1016/j.rsma.2020.101307

Lejzerowicz F, Esling P, Pillet L, Wilding TA, Black KD, Pawlowski J (2015) High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Scientific Reports 5: 13932. https://doi.org/10.1038/srep13932

Leray M, Knowlton N (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112(7): 2076–2081. https://doi.org/10.1073/pnas.1424997112

Leray M, Knowlton N (2016) Censusing marine eukaryotic diversity in the twenty-first century. Philosophical Transactions of the Royal Society B 371(1702): 20150331. https://doi.org/10.1098/rstb.2015.0331

Leray M, Knowlton N (2017) Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ 5: e3006. https://doi.org/10.7717/peerj.3006

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. Frontiers in Zoology 10(1): 1–14. https://doi.org/10.1186/1742-9994-10-34

Lobo J, Costa PM, Teixeira MA, Ferreira MS, Costa MHC, Costa FO (2013) Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. BMC Ecology 13(1): 34. https://doi.org/10.1186/1472-6785-13-34

Lobo J, Shokralla S, Costa MH, Hajibabaei M, Costa FO (2017) DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. Scientific Reports 7: 15618. https://doi.org/10.1038/s41598-017-15823-6

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2019) cluster: Cluster analysis basics and extensions. R package version 2.1.0.

McGee KM, Robinson CV, Hajibabaei M (2019) Gaps in DNA-based biomonitoring across the globe. Frontiers in Ecology and Evolution 7: 337. https://doi.org/10.3389/fevo.2019.00337

Mugnai F, Meglécz E, Costantini F, Abbiati M, Bavestrello G, Bertasi F, Bo M, Capa M, Chenuil A, Colangelo MA, de Clerck O, Gutiérrez JM, Lattanzi L, Leduc M, Martin D, Matterson KO, Mikac B, Plaisance L, Ponti M, Riesgo A, Rossi V, Turicchia E, Waeschenbach A, Wangensteen OS (2021) Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? Global Ecology and Conservation 32: e01909. https://doi.org/10.1016/j.gecco.2021.e01909

Nichols RV, Vollmers C, Newsom LA, Wang Y, Heintzman PD, Leighton M, Green RE, Shapiro B (2018) Minimizing polymerase biases in metabarcoding. Molecular Ecology Resources 18(5): 927–939. https://doi.org/10.1111/1755-0998.12895

Piñol J, Mir G, Gomez-Polo P, Agusti N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular Ecology Resources 15: 819–830. https://doi.org/10.1111/1755-0998.12355

Piñol J, Senar MA, Symondson WOC (2019) The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology 28: 407–419. https://doi.org/10.1111/mec.14776

Porter TM, Hajibabaei M (2020) Putting COI metabarcoding in context: the utility of Exact Sequence Variants (ESVs) in biodiversity analysis. Frontiers in Ecology and Evolution 8: 248. https://doi.org/10.3389/fevo.2020.00248

Prego R, Fraga F (1992) A simple model to calculate the residual flows in a Spanish ria. Hydrographic consequences in the Ria of Vigo. Estuarine Coastal and Shelf Science 34: 603–615. https://doi.org/10.1016/S0272-7714(05)80065-4

Pruesse E, Peplies J, Glöckner FO (2012) SINA: accurate high throughput multiple sequence alignment of ribosomal rna genes. Bioinformatics 28: 1823–1829. https://doi.org/10.1093/bioinformatics/bts252

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013) The SILVA ribossomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(D1): D590–D596. https://doi.org/10.1093/nar/gks1219

R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Radulovici AE, Vieira PE, Duarte S, Teixeira MAL, Borges LMS, Deagle B, Majaneva S, Redmond N, Schultz JA, Costa FO (2021) Revision and annotation of DNA barcode records for marine invertebrates: report of the 8th iBOL conference hackathon. Metabarcoding & Metagenomics 5: 207–217. https://doi.org/10.3897/mbmg.5.67862

Ransome E, Geller JB, Timmers M, Leray M, Mahardini A, Sembiring A, Collins AG, Meyer CP (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLoS ONE 12: e0175066. https://doi.org/10.1371/journal.pone.0175066

Ratnasingham S (2019) mBRAVE: The Multiplex Barcode Research and Visualization Environment. Biodiversity Information Science and Standards 3: e37986. https://doi.org/10.3897/biss.3.37986

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rognes T, Flouri T, Nichols B, Quince C, Mahe F (2016) Vsearch: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, van Horn DJ, Weber CF (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75(23): 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6): 863–864. https://doi.org/10.1093/bioinformatics/btr026

Schubert M, Lindgreen S, Orlando L (2016) AdapterRemoval v2: rapid adapter trimming, identification, and read merging Findings Background. BMC Research Notes 9(1): 1–7. https://doi.org/10.1186/s13104-016-1900-2

Shokralla S, Porter TM, Gibson JF, Dobosz R, Janzen DH, Hallwachs W, Golding GB, Hajibabaei M (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5: 9687. https://doi.org/10.1038/srep09687

Steinke D, deWaard SL, Sones JE, Ivanova NV, Prosser SWJ, Perez K, Braukmann TWA, Milton M, Zakharov EV, deWaard JR, Ratnasingham S, Hebert PDN (2021) Message in a bottle–Metabarcoding Enables Biodiversity Comparisons Across Ecoregions. bioRxiv. https://doi.org/10.1101/2021.07.05.451165

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular Ecology 19: 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Tang CQ, Leasi F, Obertegger U, Kieneke A, Barraclough TG, Fontaneto D (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences 109(40): 16208–16212. https://doi.org/10.1073/pnas.1209160109

van der Loos LM, Nijland R (2020) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology 30: 3270–3288. https://doi.org/10.1111/mec.15592

Veiga P, Torres AC, Aneiros F, Sousa-Pinto I, Troncoso JS, Rubal M (2016) Consistent patterns of variation in macrobenthic assemblages and environmental variables over multiple spatial scales using taxonomic and functional approaches. Marine Environmental Research 120: 191–201. https://doi.org/10.1016/j.marenvres.2016.08.011

Vieira PE, Lavrador AS, Parente MI, Parretti P, Costa AC, Costa FO, Duarte S (2021) Gaps in DNA sequence libraries for Macaronesian marine macroinvertebrates imply decades till completion and robust monitoring. Diversity and Distributions 27: 2003–2015. https://doi.org/10.1111/ddi.13305

von Ammon U, Wood SA, Laroche O, Zaiko A, Tait L, Lavery S, Inglis GJ, Pochon X (2018) Combining morpho-taxonomy and metabarcoding detection of non-indigenous marine pests in biofouling communities. Scientific Reports 8(1): 1–11. https://doi.org/10.1038/s41598-018-34541-1

Wangensteen OS, Palacín C, Guardiola M, Turon X (2018) DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. PeerJ 6: e4705. https://doi.org/10.7717/peerj.4705

Weigand H, Beermann AJ, Ciampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Ciamporová-Zaťovicová Z, Ferreira S, Dijkstra KB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Lesse F, Macher JN, Mamos T, Paz G, Pesic V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

WoRMS Editorial Board (2021) World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ. http://doi.org/10.14284/170

Zhang GK, Chain FJ, Abbott CL, Cristescu ME (2018) Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. Evolutionary Applications 11(10): 1901–1914. https://doi.org/10.1111/eva.12694

Supplementary material 1
Table S1, Figures S1–S5

Author: Barbara R. Leite, Pedro E. Vieira, Jesús S. Troncoso, Filipe O. Costa

Data type: zip. archiv (docx. file with descriptions and image files)

Explanation note: **Table S1.** Primer pairs used to test the efficiency of COI-5P to amplify and assess marine macroinvertebrate species in the preliminary study performed. **Figure S1.** Sampling set-up: substrates suspended horizontally and deployed in December 2016 at Toralla Island (NW Iberian Peninsula). **Figure S2.** Number of marine macroinvertebrate species detected in each substrate and sampling time by each of the four primer pairs used in the first screening of primer performance. **Figure S3.** Shared and unique marine macroinvertebrate species detected by the four COI-5P primer pairs. **Figure S4.** Taxonomic composition of marine macroinvertebrate communities for each primer pair in each substrate type and sampling time. **Figure S5.** Optimal number of clusters determined by silhouette and elbow analysis retrieved from k-means.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.5.70063.suppl1

Supplementary material 2
Tables S2, S3

Author: Barbara R. Leite, Pedro E. Vieira, Jesús S. Troncoso, Filipe O. Costa

Data type: Taxonomic classification, Species occurence

Explanation note: **Table S2.** Number of sequences (raw, usable and submitted to species level taxonomic assignment) for each primer-pair (mlCOIintF/LoboR1; LCO1490/Ill_C_R; TAReuk454FWD1/TAReukREV3), in the three substrates (slate, PVC and granite) and sampling times (3, 7, 10 and 15 months). **Table S3.** Summary of range of similarity, samples detected and notes on species assignments accuracy for the species detected using the three primer pairs (mlCOIintF/LoboR1, LCO1490/Ill_C_R, TAReuk454FWD1/TAReukREV3).

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.5.70063.suppl2